Emotion Recognition in the Wild from Videos using Images

Sarah Adel Bargal[*]
Boston University
Dept. Computer Science
Boston, MA 02215, USA
sbargal@bu.edu

Emad Barsoum
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
ebarsoum@microsoft.com

Cristian Canton Ferrer
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
cristian.canton@microsoft.com

Cha Zhang
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
chazhang@microsoft.com

ABSTRACT

This paper presents the implementation details of the proposed solution to the Emotion Recognition in the Wild 2016 Challenge, in the category of video-based emotion recognition. The proposed approach takes the video stream from the audio-video trimmed clips provided by the challenge as input and produces the emotion label corresponding to this video sequence. This output is encoded as one out of seven classes: the six basic emotions (Anger, Disgust, Fear, Happiness, Sad, Surprise) and Neutral. Overall, the system consists of several pipelined modules: face detection, image preprocessing, deep feature extraction, feature encoding and, finally, an SVM classification.

This system achieves 59.42% validation accuracy, surpassing the competition baseline of 38.81%. With regard to test data, our system achieves 56.66% recognition rate, also improving the competition baseline of 40.47%.

CCS Concepts

•**Computing methodologies** → *Activity recognition and understanding;*

Keywords

Emotion recognition; Classification; Deep features; Basic emotions; Convolutional Neural Networks; Support Vector Machines; EmotiW 2016 Challenge

1. INTRODUCTION

Automated recognition of emotions communicated through a subject's facial expression has an interesting and wide variety of applications. Emotion-aware systems would greatly empower human computer interaction. It would effectively be activating a paralyzed limb in the worlds of online courses and gaming. It has already been revolutionizing computer vision and psychology research through applications like monitoring driver state (*e.g.* fatigue state) [12], detecting depression in individuals, and diagnosis of developmental disorders of children through monitoring their facial expression and gaze during social interactions [19]. Emotion recognition from video has also been revolutionizing marketing for quantifying ad preferences [17].

Automated recognition of emotions has been addressed using different modalities: audio, video, physiological measurements, and their combinations [25]. In this work we focus on the video modality.

Emotion recognition has been addressed in terms of overall facial emotions or movement of facial muscles (Action Units) [9]. In this work, we focus on classifying emotions into the six basic emotions (Anger, Disgust, Fear, Happiness, Sad, Surprise) plus Neutral, matching the requirements of the 2016 Emotion Recognition in the Wild Challenge (EmotiW'16) [1].

There is a rich literature of hand-crafted features extracted from images and videos for encoding facial AUs and emotions. Examples include feature point tracking, dense optical flow [10], and texture-based features like Local Binary Patterns (LBP) [20, 2, 26]. These hand-crafted features are then used to train various classifiers; examples include spatial classification using Support Vector Machines, and temporal classification using Dynamic Bayesian Networks [23].

Since 2012, when the AlexNet [15] was used for image classification of ImageNet [5], deep neural networks became state-of-the-art for many vision tasks. Examples include hand gesture recognition [18], and action recognition [13, 21, 24]. Deep neural networks and deep features have also been used in emotion recognition [14]. In this work we will be using deep features to classify the videos.

Emotion recognition from videos has been addressed using static frames, and has been also addressed using ordered sequences of video frames [8]. Images have been used to aid video recognition because they are easier to collect and annotate [16]. In this work we use a spatial approach to video classification where unordered frames of the video are used, together with crowd-labeled web images.

[*]This work has been done while Sarah was an intern at Microsoft Research, Redmond.

ICMI'16, November 12–16, 2016, Tokyo, Japan
ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2997627

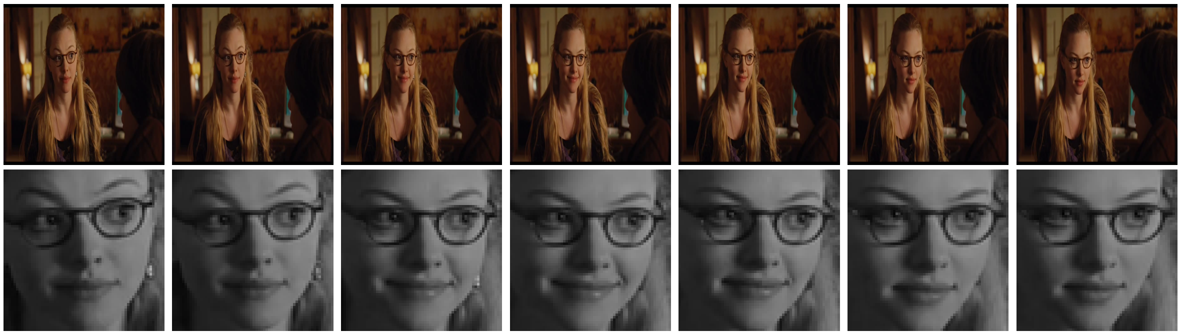

Figure 1: Sample video sequence with a ground truth label of *Happy* for the associated emotion. The frames of this video are sampled uniformly, the time dimension from left to right. Top: The original video frame, and Bottom: The pre-processed face.

2. DATASETS

In this section we present details of the EmotiW 2016 dataset and our dataset.

2.1 EmotiW'16 Dataset

The Acted Facial Expressions in the Wild (AFEW) 6.0 Dataset [7] is a dataset that consists of 1.4K trimmed video clips from movies. Being collected from movies, they are more realistic and have more challenging conditions compared to videos of facial actions deliberately produced and captured in lab conditions. "Increasing evidence suggests that deliberate behavior differs in visual appearance audio profile, and timing from spontaneously occurring behavior" [25]. The dataset is divided into training, validation, and testing. To further examine spontaneous/realistic behavior, EmotiW'16 include reality TV clips which are assumed to be more realistic than movies. AFEW 6.0 is richly annotated by human annotators for pose, character age, actor name, actor age, expression (6 basic + Neutral), gender. In this work we only use the expression annotation. A sample video sequence from AFEW 6.0 is shown in Figure 1.

2.2 Additional Dataset

We collected our own emotion image data set by crawling web images with various emotional keywords. The raw image set has over 4.5 million images. However, the majority of these images are either neutral or happy. We progressively selected around 148K images for tagging, with the latter batches focusing more and more on rare emotions. Each image was annotated by 12-15 crowd workers into one of seven basic emotions (in addition to the 6 basic emotions that were mentioned earlier, we added contempt as the seventh emotion). The numbers of images per emotion category are summarized in Table 1.

3. METHOD

We use the video modality from the provided video-audio trimmed clips provided by the EmotiW'16 challenge. We do not use other modalities like audio, and we do not use any of the provided computed features. Our system consists of a face detection module, a pre-processing module, a deep feature extractor module, a feature encoding module, and finally an SVM classification module. Figure 2 summarizes the pipeline used to obtain our results.

Table 1: Emotion category distribution of our data set

	Train	Valid	Test
neu	55180	1151	4396
hap	26271	904	1801
sur	15421	422	725
sad	11221	418	308
ang	14063	305	843
dis	3372	19	87
fea	5442	92	198
con	5329	24	26

3.1 Pre-processing

We used the face detection approach of Chen *et al.* [4]. We then crop the frame to the largest face detection. We re-size the cropped face image to match the input size of our Convolutional Neural Networks (CNNs). We then convert image to grayscale, and perform histogram equalization.

3.2 CNN Training

We train three networks: a modified VGG (13 layers) [3] based on [22], a second VGG (16 layers) [22], and RESNET (91 layers) [11]. Each of these networks is trained on the the combination of our dataset (Section 2.2) and a set of sampled frames from the AFEW training set. We follow the Probabilistic Label Drawing training process recommended by Barsoum *et al.* [3] where a random emotion tag is drawn from the crowd-sourced label distribution of an image and used as the ground truth for that image in a certain epoch.

3.3 Encoding of Deep Features

We then compute deep features using our learnt CNN models that were trained on images. We use the fully connected layer 5 (*fc5*) from the VGG13 network, the fully connected layer 7 (*fc7*) from the VGG16 network, and the global pooling layer (*pool*) from the RESNET network. For each video frame, we compute these three features: 1024-D *fc5* of VGG13, 1024-D *fc7* of VGG16, 256-D *pool* of RESNET. We normalize each of these features separately using Signed Square Root (*SSR*) and *l2* normalization. We experiment by concatenating various combinations of these features to represent a video frame.

Given the set of feature vectors representing the set of

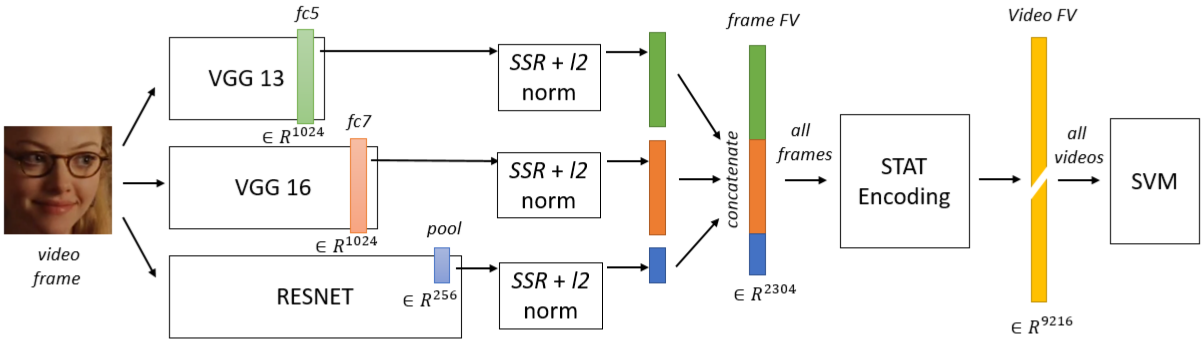

Figure 2: A depiction of the pipeline of our system. This depiction is specific to the combination of features that gave us best emotion recognition results: *fc5* of VGG13 + *fc7* of VGG16 + *pool* of RESNET. Each of there features is normalized using Signed Square Root(*SSR*) and *l2* normalization. The three normalized feature vectors are concatenated to create a single feature vector (FV) that describes this input frame. This is done for all frames of the video and inserted into the Statistical (STAT) encoding module which produces a single feature vector representing the video. This feature is then used for SVM training or classification.

video frames, we encode these features into a feature vector that represents the entire video sequence. This is done by computing and concatenating the mean, the variance, the minimum, and the maximum of feature dimensions over all video frames. This multiplies the dimensionality of the original feature vector by 4. We now normalize this encoded feature and use it for classification.

3.4 Emotion Classification

Encoded features computed as explained in section 3.3 are used to train a Support Vector Machine (SVM) to label each encoding with one of the 7 emotion classes. A One-*vs*-rest linear SVM is trained for classification using a grid search over the C paramter using 5-fold cross-validation. Best results were observed in the range $C \in [0.5, 2]$. Therefore, all results reported here are using $C = 1$. We use sklearn's LinearSVC implementation that is based on liblinear. At test time, we compute the encoded features in the same way, and use the SVM class predictions as our submission.

4. RESULTS

In this section, we present our experimental results. Table 2 shows the recognition accuracy on the use of different feature combinations on the validation data. The baseline performance provided for the sub-challenge is based on computing LBP-TOP descriptor [26] and SVR classification. We find that concatenating features from different networks gives better performance than concatenating features computed from the same network. Effectively, this is a form of regularization. We also experiment with using only the middle 90% of frames from a video sequence as we manually observe the periphery frames not being very important as they usually do not contain peak/labeled expressions (see Figure 1), but in practice, using all frames performed slightly better as demonstrated in Table 2. Table 3 presents the submission that performs best on both validation and test data.

Table 4 shows the confusion matrix of our classifier on the AFEW 6.0 validation set. It can be seen that our classifier performs well on neutral, happy and angry. However, the performance on surprise, disgust and fear is rather poor, mostly due to relatively fewer training examples.

Table 2: Validation set accuracy. ** indicates these features were computed for the middle 90% of frames of each video sequence.

Approach	Validation Acc (%)
challenge baseline [6]	38.81
op VGG13	57.07
op VGG16	55.24
op RESNET	53.66
op VGG13+*op* VGG16+*op* RESNET	57.33
fc5 VGG13	58.9
fc7 VGG16	56.02
pool RESNET	52.62
fc5 VGG13 + *fc7* VGG16 + *pool* RESNET	**59.42**
fc5 VGG13 + *fc7* VGG16 + *pool* RESNET **	59.16

Table 3: Accuracy of best submission

Approach	Validation Acc (%)	Test Acc (%)
challenge baseline [6]	38.81	40.47
fc5 VGG13 + *fc7* VGG16 + *pool* RESNET	**59.42**	**56.66**

Table 4: Confusion matrix of best-performing classifier (%).

Neu	Hap	Sur	Sad	Ang	Dis	Fea
77.78	7.94	1.59	1.59	9.52	1.59	0.00
0.00	90.48	0.00	1.59	4.76	3.17	0.00
34.78	4.35	32.61	4.35	13.04	2.17	8.70
25.00	3.33	0.00	61.67	5.00	3.33	1.67
7.81	0.00	4.69	0.00	84.38	0.00	3.12
30.00	15.00	0.00	7.50	20.00	22.50	5.00
21.74	4.35	6.52	30.43	23.91	0.00	13.04

5. CONCLUSIONS

We presented the algorithm for our submission to the EmotiW'16 challenge. The algorithm focuses on the video modality only, and achieved signficant improvement over the baseline algorithm. Future work includes leveraging the temporal relationship between video frames, and combining video emotion recognition with audio emotion recognition.

6. REFERENCES

[1] The fourth Emotion Recognition in the Wild (EmotiW) 2016 Challenge. https://sites.google.com/site/emotiw2016/.

[2] S. A. Bargal, R. el Kaliouby, A. Goneid, and A. Nayfeh. Classification of mouth action units using local binary patterns. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, 2012.

[3] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, 2016.

[4] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. *arXiv preprint arXiv:1607.05477*, 2016.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the ACM International Conference on Multimodal Interaction*, 2016.

[7] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. 2012.

[8] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 467–474. ACM, 2015.

[9] P. Ekman and W. V. Friesen. Facial action coding system. 1977.

[10] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[12] Q. Ji, Z. Zhu, and P. Lan. Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE transactions on vehicular technology*, 53(4):1052–1068, 2004.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[14] P. Khorrami, T. L. Paine, K. Brady, C. Dagli, and T. S. Huang. How deep neural networks can improve emotion recognition on video data. *arXiv preprint arXiv:1602.07377*, 2016.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] S. Ma, S. A. Bargal, J. Zhang, L. Sigal, and S. Sclaroff. Do less and achieve more: Training CNNs for action recognition utilizing action images from the web. *arXiv preprint arXiv:1512.07155*, 2015.

[17] D. McDuff, R. El Kaliouby, T. Senechal, D. Demirdjian, and R. Picard. Automatic measurement of ad preferences from facial responses gathered over the internet. *Image and Vision Computing*, 32(10):630–640, 2014.

[18] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4207–4215, 2016.

[19] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, et al. Decoding children's social behavior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3414–3421, 2013.

[20] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[21] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic. Variable-state latent conditional random fields for facial expression recognition and action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[24] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702, 2015.

[25] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.

[26] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.